chem-bla-ics

ChemSpider: the SuSE GNU/Linux of chemical databases?

Egon Willighagen

Published October 16, 2007

Citation
Willighagen, E. (2007, October 16). ChemSpider: the SuSE GNU/Linux of chemical databases?. *Chem-bla-ics*. https://doi.org/10.59350/fvykg-vc55

Keywords
Chemspider

Abstract
A molecular structure without any properties in meaningless. Structure generators can easily build up a database of molecules of unlimited size. 30 million in CAS, 20 million in ChemSpider or 15 million in PubChem is nothing yet. The value comes in when linking those structures with experimental properties.

A molecular structure without any properties in meaningless. Structure generators can easily build up a database of molecules of unlimited size. 30 million in CAS, 20 million in ChemSpider or 15 million in PubChem is nothing yet. The value comes in when linking those structures with experimental properties.

Now, chemical industry, academia and publishers have done there best in the past 50 years to maintain such databases, and decided that a commercial model was the best option to maintain such databases. This was true 50 years ago, but no longer is. ICT has progressed so much that a 20M database can be stored on a local hard disc, or site repository anyway. Moreover, and more importantly, creating a database like this is much cheaper now. These ICT developments threaten the stone age chemical databases around now. Current approaches can easily build cheap and Open chemical databases; if we only all wanted.

ChemSpider is attempting to set up the largest free chemical database, by mixing both Open data, as well as proprietary data. As such, they are attempting to achieve what SuSE and other commercial GNU/Linux distributions are trying to do: create a valuable product by complementing Open data with proprietary data when that adds value. That is, I think they are doing this. SuSE, for example, includes proprietary video drivers. ChemSpider, for example, contains proprietary molecular properties computed by ACD/Labs software (BTW, some of which can be done with Open tools too, as I will show shortly.)

Now, this poses quite a challenge: different licenses, different copyright holders, requirements to provide access to the source (for the Open data), etc, all in one system. Quite a challenge indeed, because ChemSpider is now required to track copyright and license information for each bit of information. GNU/Linux distributions do this by using a package (.deb, .rpm) approach. And, the sheer size of the database poses strong requirements if people start downloading the whole lot.

ChemSpider has had their share of critique, but the are learning, and trying to find to set up a sustainable environment for what they want to do. That might involve a revenue stream from clients if there is no governmental organization, academic institute or some society stepping in to provide financial means. A valid question would be why the did not set up a non-profit organization. But neither did SuSE, RedHat and Mandriva, but that has not stopped those from contribution to Open source.

I have no idea where ChemSpider will end up (consider that a request for a copy of the full set of Open Data), but am happy to help them distribute Open data, and even help them replace proprietary bits with open equivalents, which I'm sure the are open too. With respect to proprietary bits the are redistributing, I understand they can only relay the ODOSOS message to the commercial partners from which they get those proprietary bits, and hope they are doing. ChemSpider has the great opportunity to show that releasing and contributing chemical data as Open Data does not conflict with a healthy self-sustainable business model.